Cedat85's automatic speech recognition system

Maria Palmerini

Cedat85, Roma, Italy
m.palmerini@cedat85.com

Abstract. The speaker independent speech recognition system built by Cedat 85 for the Italian language is presented in what follows. At the beginning, the system has been tailored and trained to be included in the parliamentary reporting process, one of the major activity of the company; then further development and adaptation activities allow the use of the system in many different fields, both as different in the addressed lexicon, and applied in different application areas like, as an example, the multimedia data mining.

Keywords: ASR, speech recognition, speaker independent, data mining, Cedat 85.

1 System description

The system has been developed within a research project led in 2008 by Cedat 85 in cooperation with the European Media Laboratory in Heidelberg. It's based on the most recent IBM VoiceTailor® technology; Cedat 85 provided the whole training process (acoustic data, text data, scripts) for spontaneous Italian language.

Some of the features of VoiceTailor system are the speaker independence, the possibility to manage spontaneous speech, to use unlimited vocabularies, to use different acoustic and language models, to manage noise and the possibility to set some parameters in order to choose different strategies with respect to accuracy or speed.

The system works in a Linux environment and can run on more processors in order to have more elaborations running in parallel.

1.1 Acoustic model

During this year Cedat 85 found many different application areas for the system; this means that we have been very busy in adaptation and training activities (additional training data selection, vocabulary updating, error analysis to tune and improve the system) and we couldn't run on time the training specifically for connected digits recognition with data provided by Evalita.

For this reason, we made the test using our acoustic model, which – at the time when the test was made – was trained with almost 200 hours of audio data, from 1600 different speakers. This audio was composed by:

- audio from institutional meetings, of national and local governments (73%)
- audio from national TV (14%)
- audio from national radio (13%)

No dialects or non native speakers were included; overlapping was eliminated from the corpus. Most of the audio was clean.

1.2 Language model

As the speech recognition task was limited to digits recognition, we didn't use a proper language model for Evalita's test. The vocabulary was composed by the ten digits and the language model was generated only with unigrams.

On the contrary, since the speech recognition project started, Cedat 85 has done some work on the language model (LM) side. After the first LM built up from political reports, we've been working to a more generic LM, collecting texts from newspapers and TV broadcasts transcriptions, in order to have both written and spoken typologies of text.

At the moment, research in LM is still ongoing. According to different applications, we keep on doing various customization activities, both in terms of updating dictionaries and collecting texts from other fields to build different LMs.

2 Applications

As we said above, the first application of the system was within the parliamentary reporting process, which was Cedat 85's main business.

In last months, thanks to some partnerships with other companies, some other possibilities of application were born as, for example, an on line transcription service[1] where the customer can upload an audio file containing speech an get the automatic transcription very quickly. The last application we are working on is the integration of our ASR system into a search engine that allows to retrieve contents within huge audio/video archives; the transcription is automatically annotated and linked to the audio, so that each subject can be retrieved to be read, listened and watched.

3 Evalita test: results obtained

Table 1. Results for clean digits.

Sentence accuracy	Unit accuracy	Units	Corr.	Err.	Del.	Ins.	Sub.
89.59% (38/365)	98.05%	2360	2333	46	5	19	22

1 See www.trascrivi.com

Table 2. Results for noisy digits.

Sentence accuracy	Unit accuracy	Units	Corr.	Err.	Del.	Ins.	Sub.
78.02% (133/605)	91.03%	4036	3710	362	255	36	71

4 Discussion about results

As we mentioned in §1.1, we couldn't train the system with acoustic data provided by Evalita; we reasonably think that using our own acoustic model instead of one properly trained affected the results. For example, the audio data used to train our acoustic model don't contain sequences of digits. We are also aware that, unfortunately, using a different acoustic model made the results not comparable to the ones obtained by other participants.

Some observations can be done anyway, looking at these results. First of all, we can notice that noise made the accuracy significantly go down being our audio training material mainly clean.

Besides, as you can see from the tables above, while in clean audio the most relevant kind of error were substitutions and the deletions were only 5 out of the 46 total errors, in the noisy context the main cause of error is deletion, meaning that when the system has to do with noisy short segments of audio what can happen is that the system audio signal management just ignore the degraded signal.

Actually, we could see in our experiments that Cedat 85's system can manage quite well noise, but we must notice that in our training set there was not a lot of noise and, when there was some, this was very different from speech, both in terms of quality and of intensity.

Different training and system parameters tune up could bring to a different behaviour of the system with completely different results. Of course this can be said about small units, where “small” means words that have a small consistence in terms of phonetical material. This is something that we found also in some experiments with our own test data.

5 Next developments

The problem of noise is one of the items we are facing now; the aim would be to be able to manage raw recordings made with low quality microphones in noisy environments. From this point of view, doing this Evalita's experiment was definitely appropriate.

Other directions we are working on are: keeping on making error analysis and making the update and tuning process as automatic as possible; we are also evaluating different customization strategies on the LM side, while acoustic adaptation for a better acoustic model is on its way.

References

1. Cedat85 official website: http://www.cedat85.com
2. On line automatic transcription service: http://www.trascrivi.com
3. European Media Laboratory website: http://www.eml-research.de